Data Security and Privacy

Topic 21: Publishing Private Histogram and Using it for Classification
Reading

• Wahbeh H. Qardaji, Weining Yang, Ninghui Li: Differentially private grids for geospatial data. ICDE 2013: 757-768

• Dong Su, Jianneng Cao, Ninghui Li, Min Lyu: PrivPfC: differentially private data publication for classification. VLDB J. 27(2): 201-223 (2018)
A histogram is a graphical representation of the distribution of numerical data.
Noisy Histograms

• A histogram is a graphical representation of the distribution of numerical data
  – a partitioning of the data domain into multiple non-overlapping bins
  – the number of data points in each bin

• By adding suitable noises, publishing histogram satisfies DP
Using Histogram to Answer Range Queries

- A range query represents a hyperrectangle in the d-dimensional domain specified by the dataset, and asks for the number of tuples that fall within the bins that are completely included in the area covered by the hyperrectangle.

```
   5  1  3  4
  8  5  8 10
  1  2  0  3
  1  1  4  2
```

```
  2  3  4  1  5  1  0  6
```

- The histogram is shown above, with the red boxes indicating the bins that are completely included in the area covered by the hyperrectangle.

- The count in each bin is enclosed in the box.

- The total count of tuples that fall within the hyperrectangle is the sum of the counts in the red boxes.

- The chart on the right shows the distribution of tuples across the bins.
Utility Metrics for Range Queries (1)

• Mean Absolute Error (MAE)
  – absolute difference between the noisy answer and the true answer

• Mean Squared Absolute Error (MSAE)
  – often easier to compute
  – MSAE is the variance of the random noise
Utility Metrics for Range Queries (2)

- Mean Relative Error (MRE)
  - impact of the same absolute error is different when the true answers are different
  - the true answer may be very small, or even 0
  - chooses a threshold $\theta$ to be used as the denominator

$$\text{relative error} = \frac{|\text{true answer} - \text{obtained answer}|}{\max(\theta, \text{true answer})}$$
Example: Geospatial Data
Example: Geospatial Data
Uniform Grid

• Partition domain into \( m \times m \) cells of equal size
• Add noise to counts of each cell to satisfy differential privacy
Measuring Utility

- Error from answering range queries
  - a query is a rectangle in the data domain
Sources of Error

1. Error from satisfying Differential Privacy (noise error)
   - Adding noise from the Laplace Distribution

\[
\text{Var} \left( \text{Lap}(1/\epsilon) \right) = \frac{2}{\epsilon^2}
\]

\[
\sum_{n} \text{Var} \left( \text{Lap}(1/\epsilon) \right) = \frac{2n}{\epsilon^2}
\]
### Sources of Error

2. Error from grid: Non-uniformity error

- Assuming the data points within each cell are uniformly distributed.
Error Minimization

• Noise error: calls for coarser partitioning

• Non-uniformity error: calls for finer partitioning

• Need to choose partition granularity to minimize the sum of the two errors
Determining Grid Size

- $m \times m$ grid. Query selects a portion $r$ of the domain.
- Standard deviation of the noise error: $\frac{\sqrt{2rm^2}}{\epsilon}$
- Standard deviation non-uniformity error: $\frac{\sqrt{rN}}{c_0 m}$
- Minimize sum of two errors

$$\arg\min_m \frac{\sqrt{2rm}}{\epsilon} + \frac{\sqrt{rN}}{mc_0}$$

$$m = \sqrt{\frac{N\epsilon}{c}}, \ c \approx 10$$
Limitation of Uniform Grid

• Uniform Grid treats all regions equally
  – If a region is *sparse*, we might *over-* partitioning the region. This increases the noise error with little reduction in the non-uniformity error.
  – if a region is very *dense*, this method might result in *under-*partitioning of the region. As a result, the non-uniformity error would be quite large.
Adaptive Grid

• Adapt the level of partitioning based on the number of data points in each region
  – If a region is dense, use finer granularity to reduce non-uniformity error
  – If a region is sparse, use a more coarse grid
Adaptive Grid
Adaptive Grid

• Two level partitioning:

1. Lay a coarse $m_1 \times m_1$ grid over the data domain and obtain a noisy count for each cell

2. Partition each cell into an $m_2 \times m_2$ grid, where $m_2$ depends on the noisy count of the cell

3. Apply constrained inference
Adaptive Grids

• Choosing Parameters \((m2)\):
  – Average noise error:
    \[
    \sqrt{\frac{(m_2)^2}{4}} \frac{\sqrt{2}}{(1-\alpha)\epsilon}
    \]
  – Average non-uniformity error:
    \[
    \frac{N'}{c_0 m_2}
    \]
  
  \[
  m_2 = \left\lceil \sqrt{\frac{N'(1-\alpha)\epsilon}{c_2}} \right\rceil
  \]
Adaptive Grids

• Choosing Parameters ($m_1$):
  – Parameter is less critical, since the second level adapts to the count of each cell
  – In general, we want it to be less than the choice for uniform grids.

$$m_1 = \max \left( 10, \frac{1}{4} \left\lceil \sqrt{\frac{N \varepsilon}{c}} \right\rceil \right).$$
PrivPfC: Private Publication for Classification [Su et al. 16]

Sensitive Dataset → "Good" Grid → + noise → Noisy Synopsis → generate Synthetic Dataset

Enumeration → Private Selection → Candidate Grid Pool

Considers DP

Classification
Generation Hierarchy & Grids

Relationship

Education-num

Level-1

Level-2

Level-3

- Wife

- Husband

In-family

Not-in-family

Unmarried

Any

1-20

1-10

10-20

1-5

5-10

10-15

15-20

Grid 1

Grid 2

In-family

Not-in-family

Unmarried
**Histogram Classifier**

Dataset → Project → Grid 1

<table>
<thead>
<tr>
<th>Relationship</th>
<th>1-10</th>
<th>10-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-family</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Not-in-family</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Unmarried</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Majority Voting
Consider the Noise Impact

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Education-num 1-10</th>
<th>Education-num 10-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-family</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Not-in-family</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Unmarried</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

+Noise

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Education-num 1-10</th>
<th>Education-num 10-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-family</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>Not-in-family</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Unmarried</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

Label is flipped
Quality Function

- Number of correctly classified points is a random variable
- Use expectation as the quality

\[ \text{qual}(g) = \sum_{c \in g} n_c^+ \cdot p_c^+ + n_c^- \cdot (1 - p_c^+) \]

- \( n_c^+ \): number of points in \( c \) with positive label
- \( p_c^+ \): prob of positive label is the majority after injecting the noise

\[ p_c^+ = \Pr[n_c^+ + Z_c^+ > n_c^- + Z_c^-] \quad Z_c^+, Z_c^- \sim \text{Lap}(1/\epsilon) \]

\( \epsilon \) is budget for perturbation
## Quality Scores under Different Settings

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Education-num</th>
<th>1-10</th>
<th>10-20</th>
<th>1-5</th>
<th>5-10</th>
<th>10-15</th>
<th>15-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-family</td>
<td></td>
<td>20</td>
<td>16</td>
<td>2</td>
<td>6</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Not-in-family</td>
<td></td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Unmarried</td>
<td></td>
<td>6</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\varepsilon = 0.2 \\
\text{qual}(g_1) = 47 \\
\text{qual}(g_2) = 48
\]
Sensitivity of Quality Function

\[ \Delta_{\text{qual}} \] vs. \( \epsilon \) (Perturbation budget)

- Sensitivity < 1.1
PrivPfC: Private Publication for Classification

Algorithm 7 PrivPfC: Differentially Privately Publishing Data for Classification

Input: dataset $D$, the set of predictor variables $A$ and their taxonomy hierarchies, total privacy budget $ε$, maximum grid pool size $Ω$.

\[ε_N \leftarrow 0.03ε, \quad ε_{sh} \leftarrow 0.37ε, \quad ε_{ph} \leftarrow 0.6ε\]

\[\hat{N} \leftarrow |D| + \text{Lap}(1/ε_N)\]

\[T \leftarrow 20\% \cdot \hat{N} \cdot ε_{ph}\]

\[H \leftarrow \text{Enumerate}(A, Ω, T)\]

Comment: privately select grid

for $i = 1 \rightarrow |H|$ do

\[q_i \leftarrow \text{qual}(H_i)\]

\[p_i \leftarrow e^{-(q_i ε_{sh})/2}\]

end for

\[h \leftarrow \text{sample } i \in [1..|H|] \text{ according to } p_i\]

Initialize $I$ to empty \quad Comment: privately perturb grid

for each cell $c \in h$ do

\[\hat{n}_c^+ \leftarrow n_c^+ + \text{Lap}(1/ε_{ph})\]

\[\hat{n}_c^- \leftarrow n_c^- + \text{Lap}(1/ε_{ph})\]

Add $(\hat{n}_c^+, \hat{n}_c^-)$ to $I$

end for

Round all counts of $I$ to their nearest non-negative integers.

return $\hat{I}$
## Summary

<table>
<thead>
<tr>
<th>Idea</th>
<th>Differentially Private Optimization for ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single Workload</td>
</tr>
<tr>
<td>Idea</td>
<td>Breaking task into query workload</td>
</tr>
</tbody>
</table>

### Pros

<table>
<thead>
<tr>
<th>Single Workload</th>
<th>Non-interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customized for a specific task</td>
<td>- Re-usable</td>
</tr>
<tr>
<td></td>
<td>- Preserve data distribution</td>
</tr>
</tbody>
</table>

### Cons

<table>
<thead>
<tr>
<th>Single Workload</th>
<th>Non-interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Over-divided privacy budget</td>
<td>- Not customized for a specific task</td>
</tr>
<tr>
<td>- No more data distribution</td>
<td></td>
</tr>
</tbody>
</table>
Experiments on DP Classification

• Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Dim</th>
<th>#Num/#Cate</th>
<th>#Records</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>15</td>
<td>6/8</td>
<td>45,222</td>
<td>&gt;50K/yr</td>
</tr>
<tr>
<td>Bank</td>
<td>21</td>
<td>10/10</td>
<td>41,188</td>
<td>Sub. Deposit</td>
</tr>
<tr>
<td>US</td>
<td>47</td>
<td>15/31</td>
<td>39,186</td>
<td>&gt;50K/yr</td>
</tr>
<tr>
<td>BR</td>
<td>43</td>
<td>14/28</td>
<td>57,333</td>
<td>&gt;300/mo</td>
</tr>
</tbody>
</table>

• Setup
  • Decision tree and SVM
  • Vary epsilon from 0.05 to 1.0
  • Misclassification rate
Decision Tree Comparison on Adult Dataset

PrivPfC does the best on Decision Tree task
SVM Comparison on Adult Dataset

PrivPfC does the best on SVM
Next Lecture

- Meanings and caveats of DP